INTERNET DOCUMENT INFORMATION FORM

A . Report Title: TCP Performance over ACTS

B. DATE Report Downloaded From the Internet 9/22/98

C. Report's Point of Contact: (Name, Organization, Address, Office Symbol, & Ph #): **Nasa Lewis Research Center**
21000 Brookpark Road
Cleveland, OH 44135-3127
ATTN: Doug Hoder (216) 433-8705

D. Currently Applicable Classification Level: Unclassified

E. Distribution Statement A: Approved for Public Release

F. The foregoing information was compiled and provided by:
DTIC-OCA, Initials: VM___ **Preparation Date**:_9/23/98_____

The foregoing information should exactly correspond to the Title, Report Number, and the Date on the accompanying report document. If there are mismatches, or other questions, contact the above OCA Representative for resolution.

TCP Performance over ACTS [1]

C.E. Fair
High Performance Netwoking Section - Scientific Computing Division
National Center for Atmospheric Research
Boulder, Colorado

ABSTRACT: *Coupled atmospheric and hydrodynamic forecast models were executed on supercomputing resources of the National Center for Atmospheric Research (NCAR) in Boulder, Colorado and the Ohio Supercomputing Center (OSC) in Columbus, Ohio respectively. The interoperation of these two models on distributed, high performance computing platforms required the transfer of large three dimensional data sets at very high information rates. High capacity, terrestrial fiber optic transmission system technologies were integrated with those of NASA's Advanced Communications Technology Satellite (ACTS) in Geosynchronous Earth Orbit (GEO) to test integration of the two systems. Operation over such a spacecraft required the modification of standard data communications protocol configurations to facilitate their ability to perform efficiently over the hybrid network. Protocol performance tuning enabled this architecture to facilitate high data rate communications between end-systems not readily accessible to high performance terrestrial fiber optic transmission systems. Obviating the performance degradation often found in contemporary earth/satellite hybrids.*

1 Introduction

Interworking of legacy supercomputer systems with the revitalized technologies embodied by the high power NASA Advanced Communications Technology Satellite (ACTS) required the integration of high performance terrestrial systems which were complementary with the spacecraft's expanded capabilities. Dissimilar physical layer technologies required integration, providing an opportunity to investigate their interoperability over high speed terrestrial and satellite media. Native High Performance Parallel Interface (HIPPI) traffic from Cray supercomputers was converted to Synchronous Optical Network (SONET) framing, the physical layer for the ACTS satellite transmisson system. Asynchronous Transfer Mode (ATM) cell streams generated by high performance workstations controlling model interaction and visualization of model output were converted to HIPPI, then back to SONET. A key component in the hybrid architecture's success was the ability to facilitate interoperation of diverse physical layer technologies simultaneously.

A critical factor in the success of the hybrid model's interoperability was the optimization of Transmission Control Protocol/Internet Protocol (TCP/IP) for the high performance satellite channel. The high data rates afforded by SONET combined with the great latency (delay) of the space segment exceeded the domain of classical TCP functionality. This is manifested in the delay for the acknowledgment of transmitted packets being great enough for the sender's transmission window to time out well before its packets have reached their destination.

This results in a very inefficient "stop and wait" state on the sender, shutting off the flow of data between the application, the operating system kernel and the transmission media until some reply by the receiver elicits an appropriate response from the sender [3].

Performance is impacted as throughput declines from the sender's inability to keep the channel full; the bit length of transmitted data being far less than that which the channel can accommodate. For example if a packet of 64 K bytes is sent over an OC-3c link (155 Mbps) with a 550 ms round trip time delay (RTT), under the stop and wait scenario the maximum throughput would be slightly less than 1 Mbps. This equates to a link utilization of 155 times *less* than an OC-3c link can efficiently support. To address this defiency, performance enhancements to the TCP Automatic Retransmission Request (ARQ) or sliding window were invoked to match the performance of TCP to that of the hybrid network. [1]

[1] This work was supported by ARPA Grant 93-1 and ARPA Order A701.

2 Physical Layer Configuration: IP Over HIPPI -> SONET

In local environments where the time-distance separation between machines is slight, HIPPI interfaces and HIPPI switches may directly perform the interconnection of end systems. HIPPI uses a 32-bit parallel architecture transmitting over a simplex channel at 800 Mbps, channel pairs establishing full duplex data communications. Wide area interoperability is facilitated by sockets to layered protocol suites such as connectionless User Datagram Protocol (UDP) or connection-oriented Transmission Control Protocol (TCP), both providing OSI layer 4 services.

Prior to transmission via the satellite, model data sets transferred by the Cray's HIPPI interfaces via a HIPPI switch were converted to a serial SONET stream by a HIPPI/SONET gateway developed at the Los Alamos National Laboratory (LANL). The HIPPI streams were framed in SONET OC-3c (155.52 Mbps) Synchronous Payload Envelopes (SPE) and transferred to the ACTS ground segment High Data Rate (HDR) terminal via Single Mode optical fiber. The SONET Section and Line Overhead were terminated by the transmitting HDR and regenerated by the receiving HDR SONET section.

3 Physical Layer Configuration: IP Over ATM -> HIPPI -> SONET

The modeling components on the NCAR Atmospheric Model, the Penn State Hydrodynamic Model and the Lake Erie Wave Model were linked using Parallel Virtual Machine (PVM) [2]. PVM is a software library which provides a uniform, parallel computing architecture independent of the underlying hardware and network topology. PVM also linked the models to data managers, software written to filter and route model output streams. This data was fed to visualization components executed on high performance Silicon Graphics Inc. (SGI) workstations. Control and evaluation of the simulation was afforded by this component. Collaboration among the scientific researchers was facilitated by packet-based (IP) over Asynchronous Transfer Mode (ATM), video conferencing to/from each site, also via ACTS.

The video conferencing and collaborative workstations were directly connected via ATM to Fore Systems ASX-200 ATM switches. IP packets segmented into 53 byte ATM cells by the workstation's ATM Network Interface Cards (NIC) were converted to HIPPI by a NetStar GigaRouter and routed through the HIPPI switch to the HIPPI/SONET gateway. As with the Cray output, this HIPPI stream was then framed in SONET OC-3c SPE and transferred to the HDR terminal over single mode fiber. SONET frames from the workstations were multiplexed with model data via the single OC-3c ACTS channel, completing the terrestrial complement of the hybrid.

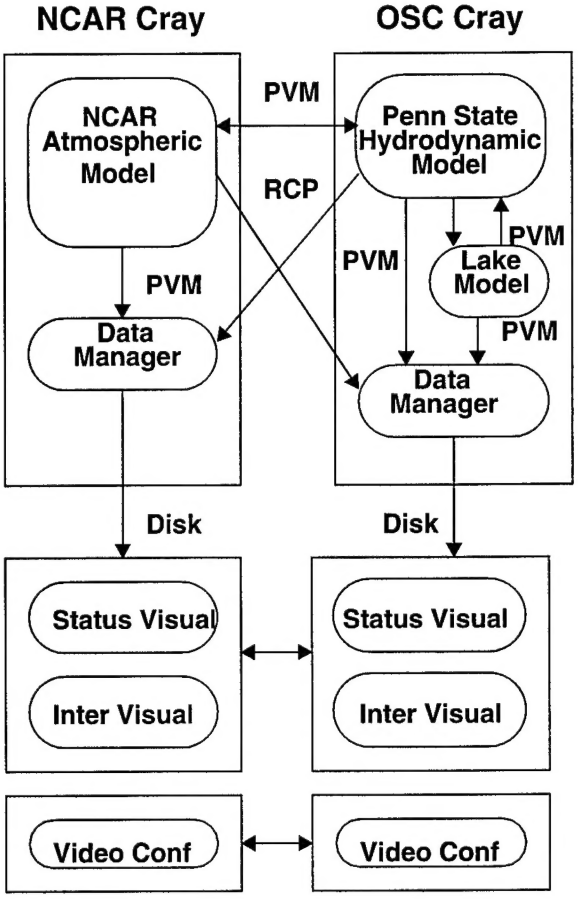

Figure 1: Modeling component interaction via PVM and interactive model control /output.

4 TCP Performance Extension

The ability of an end system to transmit data is ultimately limited by the information rate capacity of

the transmission medium. Efficient use of the medium is achieved by maintaining transmission rates at or close to the maximum. The combination of this data rate capability and the Round Trip Time (RTT) between source and destination specifies how much data can flow at any instant between the sender and receiver.

TCP is a reliable, end to end connection-oriented transport layer protocol which uses a sliding window based flow control system or ARQ to recover from loss or corruption of data over the medium. To achieve this, TCP requires the source to hold the data transmitted in buffer for a minimum of the time required to send the data to the destination and receive an acknowledgment from the receiver or the RTT (link delay or latency). Should data be corrupted or lost the *entire* contents of the source buffer is then retransmitted by the sender during the recovery process [3].

Maximum performance is obtained from TCP not just from high information rates but from the product of the information rate and the RTT. This "*bandwidth-delay*" product (BDP) [1] is equivalent to the amount of unacknowledged data outstanding at any instant on the transmission medium. The bandwidth-delay product then corresponds to the minimum buffer size or window size which will keep the "pipe"or link full and provide expeditious recovery to congestion or loss. The larger the window, the more data can be outstanding and the capacity of the data link maintained at or near its maximum capacity.

TCP window size corresponds to the size of the socket buffer space or send and receive buffers in both the source and destination UNIX operating system kernels. During connection establishment the source and destination negotiate the size of this window, facilitating the smooth, continuous flow of data for the duration of the connection. To provide efficient use of high capacity links with high latency, very large window sizes are required.

In the original TCP specification, RFC 793 [3], the TCP header contains a 16 bit window size field which corresponds to the receiver's window size. The 16 bit field can support a maximum window size of 2×10^{16}, 64 K Bytes. RFC 1323 [1] prescribes a window scalability option for the TCP header which can accommodate larger window sizes, up to 1 Gbyte. This option can improve the performance of modern high capacity networks with inherent high bandwidth-delay products. The extension maps the standard 16 bit window size field to a 32 bit value and uses the window scale option to bit-shift this value, producing a new maximum window size value.

The window scale option occupies 3 bytes in the TCP header, it specifies the type of option as window scale and the second 3 bytes the length of the option and the shift count. The window scale indicates the sender is able to accept send and receive buffer or window scaling and sends the scale factor to the receiver. The window scale is a log $base_2$ value and the shift count is the number of bits the receiver's window value is to be right shifted. Right shift applies to the default TCP window specified in the TCP header [3]. Values less than the 2×10^{16} maximum will only be right shifted by the shift factor.

An application may set a larger window size with the *setsockopt* call, based on the available buffer space of the operating system kernel. The implementation of window scale will then determine the appropriate shift factor. The maximum window shift can be obtained starting with a default maximum window size of 2×10^{16} and a scale factor of 14, resulting in a maximum window size of approximately 1 GByte,
$(2 \times 10^{16} * 2 \times 10^{14} = 2 \times 10^{30} = 1.073$ Gbyte).

The RTT of the data link and the maximum information rate must be known to facilitate performance enhancements using the window scale option to adjust window size. Application of window scale options requires that the operating system kernels of both sender and receiver accommodate the extensions to TCP performance outlined in RFC 1323 [1].

The maximum amount of socket buffer space available to the operating system kernel must be great enough to accommodate the window scale factor anticipated. This often entails kernel reconfiguration requiring *all* of the available memory of a shared memory, multiprocessor machine; effectively dedicating the entire machine to a single application and perhaps even a single user.

5 Integration and Test Configurations

Various levels of system integration were performed. Commensurate continuity and performance tests were made to validate progress and functionality of the physical layer integration and the TCP performance enhancements prior to advancement to the next level.

The levels were:

- Earth Station Installation
- Network hardware installation
- Window scale optimization studies
- Satellite loopback testing
- End-to-end connectivity tests via ACTS

As all physical layer data streams were converted to SONET for transmission over the ACTS spacecraft, interoperation of the two HIPPI/SONET gateways was verified at NCAR prior to the shipment and installation of the second gateway at the OSC site. Various configurations of both HIPPI/SONET gateways were tested in loopback with single Cray connectivity to test continuity and validate raw HIPPI and HIPPI/SONET performance for each. The gateways were then tested between two local Crays, an EL-92 and the J-916 which would execute the NCAR atmospheric model during the experiment. This was to simulate end-to-end connectivity involving both gateways and different machines. Finally each gateway was tested between two local Crays via a local loopback at the NCAR HDR. HIPPI performance was validated by a simple program which writes 10 Mbyte buffers of raw HIPPI data across logical interfaces on a NSC PS-32 HIPPI switch to a single or pair of Crays [4].

HIPPI performance tests were made for the following HIPPI/SONET gateway configurations:

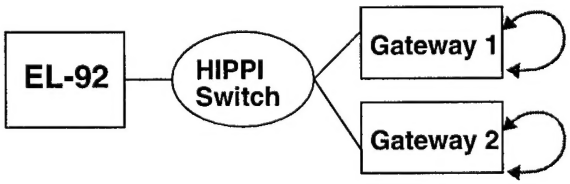

Figure 2: Single Cray looped through each gateway (individually) via HIPPI switch.

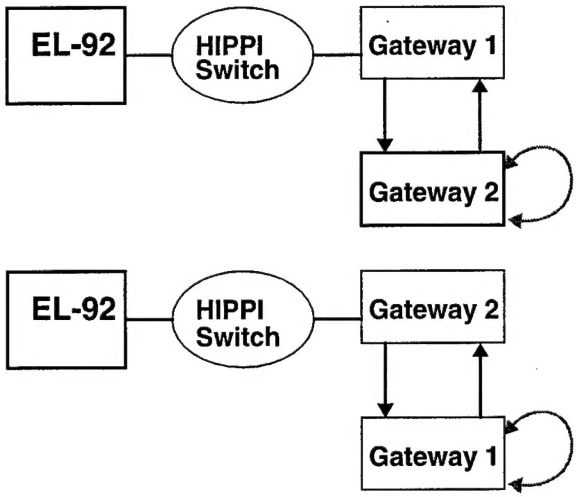

Figure 3: Single Cray looped through both gateways (back-to-back) via HIPPI switch.

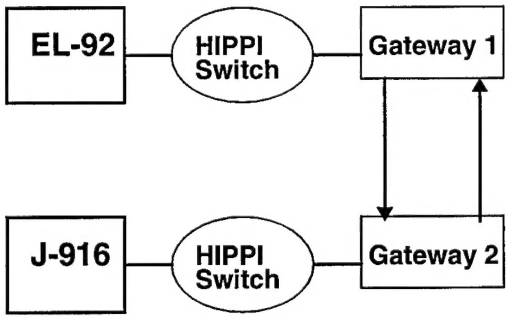

Figure 4: Cray to Cray via both gateways (back-to-back in series) via HIPPI switch.

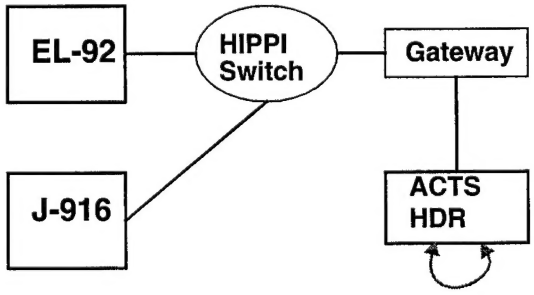

Figure 5: Cray to Cray via HIPPI switch, single gateway and ACTS HDR terminal in loopback.

TCP performance tests were conducted between two NCAR Crays interconnected similarly to the above HIPPI test configurations and via ACTS:

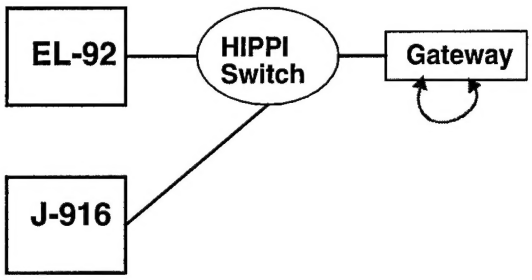

Figure 6: Cray to Cray via single gateway in loopback via HIPPI switch.

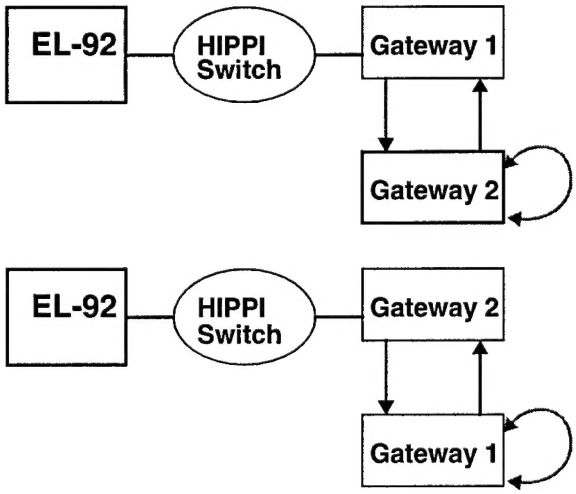

Figure 7: Cray to Cray via both gateways (back-to-back) via HIPPI switch.

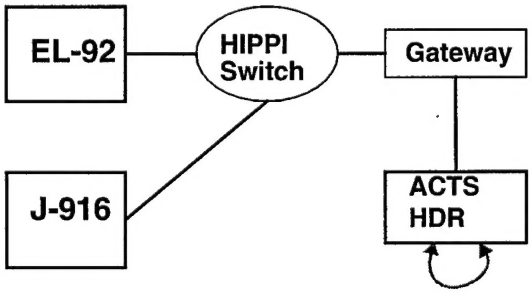

Figure 8: Cray to Cray via HIPPI switch, single gateway and ACTS HDR terminal in loopback.

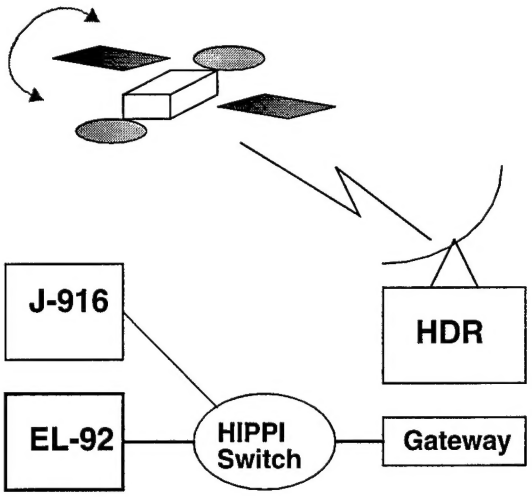

Figure 9: Cray to Cray via HIPPI switch, single gateway over ACTS OC-3c in loopback.

Once validated latency for the round trip spacecraft OC-3c channel was incorporated into a Long Link Emulator (LLE) to simulate the satellite delay for performance enhancement and application development when spacecraft time was not available.

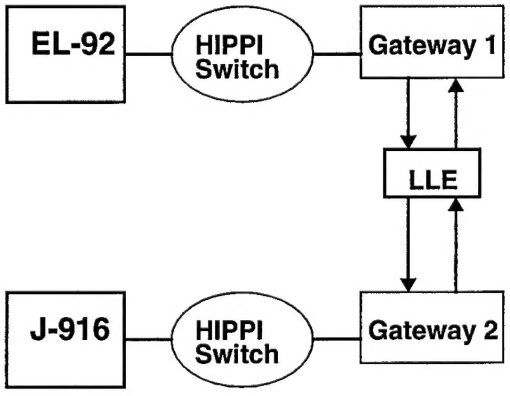

Figure 10: Cray to Cray via both gateways and Long Link Emulator (LLE) via HIPPI switch.

Connectivity between NCAR and OSC was established only after all configurations were successfully tested. NCAR to OSC connectivity was made through a configuration at each site similar to that in Figure 9 with the spacecraft MSM configured for straight through (end-to end) connectivity.

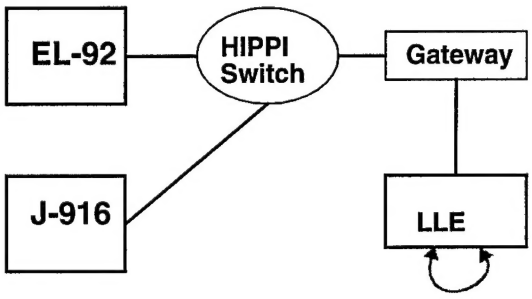

Figure 11: Cray to Cray via single gateway, Long Link Emulator in loopback (LLE) via HIPPI switch.

6 TCP Performance Tuning

The required TCP window size for the Crays and hence the window shift was determined from the bandwidth-delay product of the hybrid. The round trip time was calculated empirically, then validated against actual spacecraft round trip time measurements.

6.1 Bandwidth - Delay Product

A round trip for a packet and its corresponding acknowledgment required *two* satellite hops. The packet traversing the link once enroute to the destination where if received correctly, it elicits an acknowledgment from the receiver. The acknowledgment then traverses the link back to the sender, completing the round trip. The calculation of the time required for this round trip between two Cray machines at NCAR interconnected via the satellite loopback (Figure 9) was made based on a one-way propagation time between the NCAR HDR and the spacecraft multiplied by two.

39° 58' 39" north latitude
105° 16' 28 " west longitude
6,113 feet above mean sea level

For the purpose of calculation, latitude was rounded to 40°, the radius of the earth was taken to be 3,960 miles and the orbital altitude of the spacecraft was assumed to be 22,300 miles above the earth's equator (the orbital longitude of the spacecraft was not considered). The speed of light, *c* was taken to be 186,400 miles/second in the vacuum of space and the atmosphere. The *Law of Sines* was used given a triangle formed by the center of the earth (point A), the earth segment antenna (point C) and the spacecraft (point B). Triangle ABC subtends the angles A, B and C respectively. The sides of triangle ABC opposite these angles are sides a,b and c.

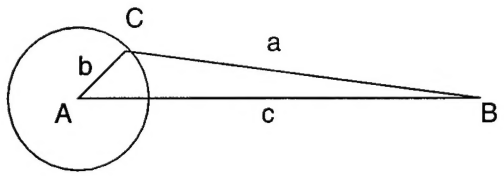

Figure 12: Triangle formed by relative position of ACTS spacecraft and the earth.

An empirical value of 502 ms was determined to be the round trip time between the NCAR and the ACTS satellite. This value provided a baseline against which all future round trip time measurements would be validated.

Figure 13 illustrates measurement of the actual RTT using a series of 64 byte Internet Control Message Protocol (ICMP) queries (ping) sent between a Cray Y-MP8 and a Cray EL-92 over the ACTS spacecraft's Microwave Switching Matrix (MSM) in loopback (bent pipe).

The average ICMP measured RTT of 541 ms was compared to the empirical value of 502 ms and an assumption was made that the additional 39 ms was due to terrestrial delay somewhere between the Crays and the NCAR HDR terminal.

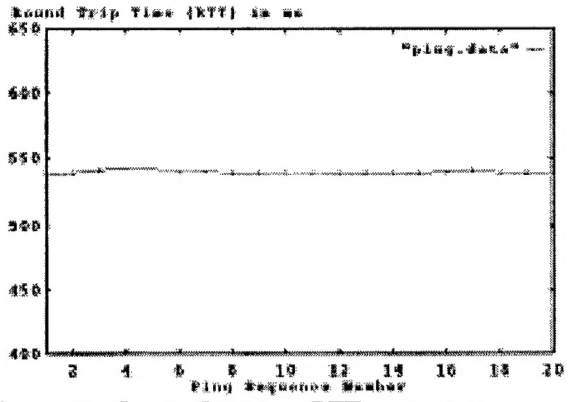

Figure 13: Space Segment RTT measurement round-trip (ms) min/avg/max = 539/541/552

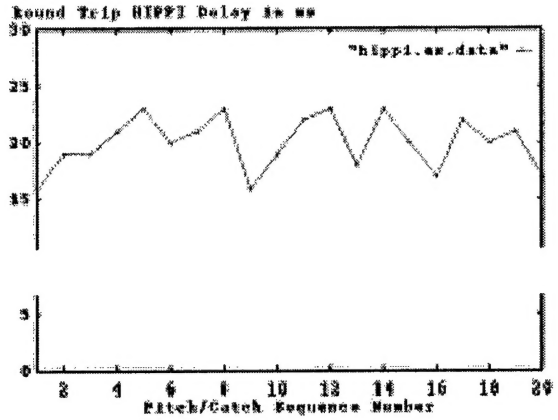

Figure 14: Round trip HIPPI delays across single HDR Earth Terminal.

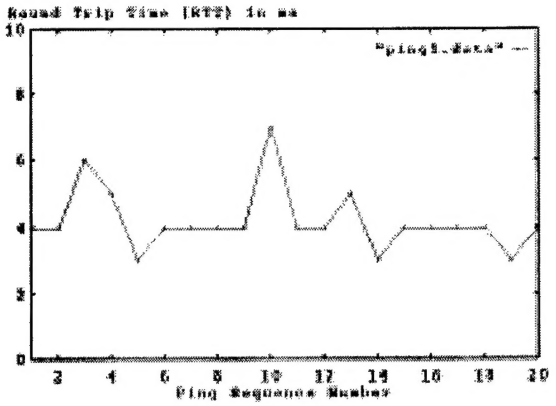

Figure 15: Terrestrial segment RTT measurement.

Raw HIPPI tests were used to determine the latency across the HDR. Blocks of data were transferred between the Crays, via the HIPPI switch, the HIPPI/SONET gateway and the HDR in a loopback configuration (Figure 8). This configuration resulted in each packet passing through the HDR twice before being "read" by the receiver; once upon receipt and again for retransmission back to the receiver. An example output of these "Pitch and Catch" [4] operations is illustrated below.

Output channel -- HXCF_HIPPI is set
HXCF_HDR is set (user buffer has FP header)
HXCF_IND is zero (I-field not in user buffer)
HXCF_ISB is zero (Short burst at end)
Data length = 65536 bytes.
FP header: 858000180000ffe0
I-field is 702a3d1
HXC_SET for output OK.
Write of 10485760 bytes completed successfully
Elapsed microsecs = 749291
Fastest single block was 122.269 Mbits/sec
Slowest single block was 40.016 Mbits/sec
Overall data rate was 111.954 Mbits/sec
Catch completed successfully
..**First packet delay was 21504 microsec.**
..Overall data rate with delay was 109.334 Mb/sec

Figure 14 shows the Roundtrip delays for the first HIPPI packet to return from the looped HDR terminal using these "pitch and Catch" tests. This data illustrates terrestrial earth station delays consistently in the 20 ms range.

Data measurements plotted in Figure 15 with the HIPPI/SONET gateway in loopback and the LLE set to emulate a 0 ms RTT latency, reveal small terrestrial delays in the 4 ms range. This suggests that the bulk of the additional delay observed in the hybrid is the result of *two* trips for each transmitted and received packet through the HDR terminals at each end of the satellite link. Approximately 20 ms for the transmitted packet in one direction and an additional 20 ms for the acknowledgement's return trip in the opposite direction. Thus the additional 35-40 ms delay observed can be attributed to delays in the terrestrial segment of the hybrid, most likely encountered in the HDR.

6.2 Bandwidth - Delay Product and TCP Window Size

The SONET OC-3c information rate is 155.52 Mbps, however Section and Line Overhead are terminated by the HDR reducing the transmitted information rate. The nettest/nettestd TCP/IP performance measurement utilities are a TCP application and thus TCP/IP overhead must be accounted for also as reducing the available information rate. A more conservative rate of 135 Mbps was selected for use in bandwidth-delay product calculations to account for the SONET and TCP/IP overhead.

(135 x 10^6 b/s) * (541 ms) = 73,035,000 bits/8 bits/byte = 9,129,375 bytes

The bandwidth-delay product above specifies the

minimum send/receive buffer or window size required for optimum TCP performance on the hybrid network. This value specifies a window shift of 8 (2×10^8). While a window size of 9,129,375 bytes is optimal, the window shift will be set for the next larger increment. A shift of 8 would yield a much larger buffer than the computed value,

$(2 \times 10^{16}) * (2 \times 10^8) = 16{,}777{,}960$ bytes

While a window shift of only 7 would not provide a window that is large enough to accommodate the bandwidth-delay product of 9,129,375 bytes.

(2 x 1016) * (2 x 107) = 8,388,480 bytes

Specification of the correct window size in the nettest utility with the -b option will set the correct buffer size. A window shift of 8 will accommodate that size and any size up to the maximum shift value of 8.

A modified version of the Cray UNICOS TCP test utility nettest and nettestd [5] was used to measure the effect of the window size and window size changes on throughput performance. The nettest/ nettestd utility performs client and server functions for measuring network throughput of interprocess communication. The nettest program establishes a connection with the nettestd program which performs the server function, waiting for the nettest client to initiate the process communications. As with any TCP connection, the window scale option is sent at connection establishment in the <SYN> segment. Thus the window scale value is fixed when the connection is made and remains for the duration of the connection. The nettest program writes a number of bytes to the nettestd program which reads a number of bytes and reports the throughput. Nettestd in turn writes a number of bytes to nettest which reads them and also reports the performance. The performance of the two processes is averaged to disclose an average data throughput rate for the TCP connection.

Modification to the nettestd source code was required because the UNICOS release at the time of the experiment nettestd did not have the capability to allow the user to set the window size or shift factor. Rather the default maximum window size defined in the operating system kernel (in this case $2 \times 10^{16} = 65{,}536$ bytes) was used by nettestd. To overcome this limitation it was put forth that perhaps during connection establishment nettestd "spoofs" the server that the size of its receive buffer is the same as the sender's, a requirement for smooth data flow in both directions. The connection is established but an imbalance in buffer sizes between the source and destination may result in asymmetric transfer rates observed between the data sent between the client and the server.

While experimenters were unable to verify that this spoofing is in fact occurring, modification of the nettestd code allowed the user to define the window shift for both nettest and nettestd sessions by defining the -s and -b options in the nettest program. Upon connection establishment the specified window sizes are set on both client and server, resulting in a symmetric data transfer between client and server assuming similar loads on the two machines.

7 Results

Performance tests using the modified nettest/ nettestd programs were executed for the various configurations. The results of the tests between different Cray platforms were validated against each other for bent pipe tests as well as full connectivity, end-to-end tests between NCAR and OSC. Tests were made using various combinations of window shift and buffer size to determine and validate optimal TCP performance parameters. These results are plotted in Figures 16 through 18.

7.1 Validation of TCP Window Shift and Window Size

Window sizes and subsequent window shift factors were calculated prior to beginning every test session by dynamic RTT measurement and BDP calculation. This was done because of slight variation found in RTT measurements between the Crays. While small, such variations are capable of significant differences in window sizes and ultimately kernel buffer requirements. Plots were taken for BDPs using the average and maximum RTT measurements such as those plotted in Figure 14 between a Cray EL-92 and a Cray J-916 at NCAR.

Figure 16 illustrates a plot of performance against window shift variations in two optimal window sizes as a result of BDP variations. It also shows the rather dramatic performance validation of the correct window shift for the given BDP and TCP window size.

The slight variations in window size for the two plots seems trivial as long as the minimum window size used was greater than or equal to the BDP. Performance seems to suffer greatly however if the window shift applied is less than that required for the window size.

To verify this, the effect of sub-optimal window sizes was explored by varying the window sizes against a fixed window shift of 8 validated previously. Figure 17 indicates that performance improves linearly with improvements in window size until the optimal window size for the BDP is reached. For these tests the BDP parameters verified in the plots of Figures 13 through 15 were used.

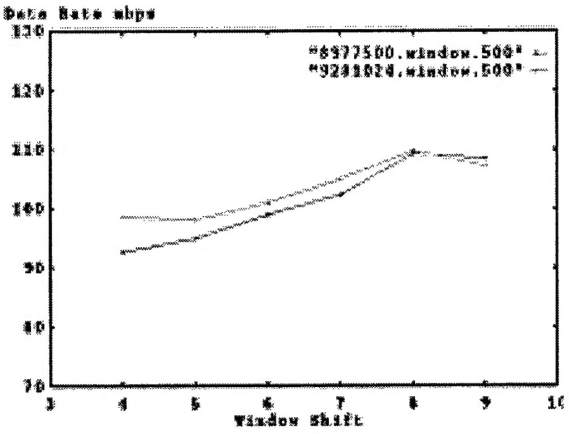

Figure 16: Window shift vs. Performance.

The results indicate that the optimal window shift and socket buffer size for the ACTS channel was a shift of 8 and a window size of no less than 9,095,625 bytes, validating the empirical results.

A TCP window size of 1 M byte and a window shift of 8 was chosen as a baseline for all further tests in loopback and end-to-end configurations.

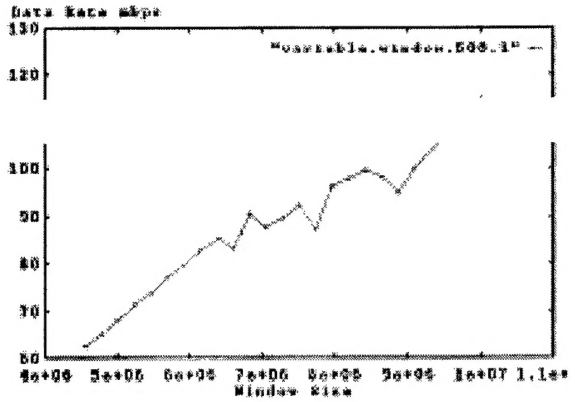

Figure 17: Window size vs. Performance

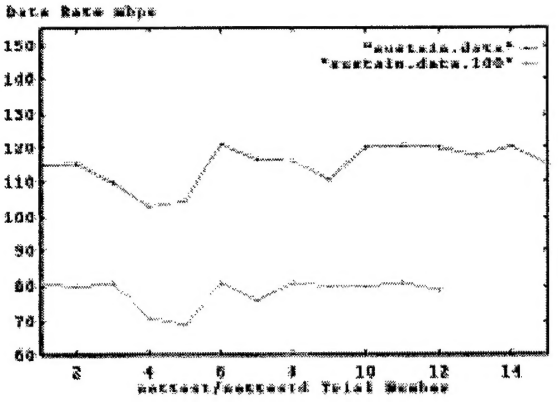

Figure 18: Effect of changes in number of nettest/ nettestd read/write operations on performance.

7.2 Factors Affecting nettest/nettestd Performance

It was thought that large amounts of data were required to keep the "pipe full" and ensure sustained link efficiency. Thus the number of read or write operations and the number bytes to be read or written with each operation was varied to affect the amount of data in the pipe at any given time.

First the number of times data was read/written was set to 100 and the number of bytes for each of the operations was 3 Mbytes each. A disparity in performance is evident between the results of these tests and tests where the number of times data was read/written was set to1,000 times and 3 Mbytes each.

It is not clear why more optimal performance is observed when the amount of data is increased by an order of magnitude. The amount of data transmitted is sufficient to "fill the pipe" in either case. While this is puzzling the total 'Real' time to effect the transfer increases by a only factor of seven, indicating more efficient use of the channel. This phenomena was observed in bent pipe configurations between machines at NCAR and in later end-to-end tests between NCAR and OSC Crays.

In spite of this anomaly, 1,000 read/write operations appeared to be the number of operations which produced optimal throughput performance. It was used for all subsequent tests.

Figure 18 illustrates that repeated nettest/nettestd suites executed using the validated window size and window shift parameters in the bent pipe configuration yielded very consistent performance in the 120 mbps range as long as the number of nettest/nettestd operations was adequate.

7.3 End-to-End Connectivity via ACTS

Having spent a fair amount of time validating Cray to Cray performance in bent pipe configurations at NCAR, nettest/nettestd validation tests on the NCAR Cray J-916 and the OSC Cray Y-MP8 were commenced. This would be the final test configuration and the one that the coupled atmospheric and hydrodynamic models at each site would run on. The objective was to use all of the parameters from the bent pipe configurations tests to validate the end-to-end performance.

It was expected that since all TCP performance parameters had been successfully validated, once end-to-end connectivity over the spacecraft was established both machines would interoperate successfully via this paradigm.

This was not to be. Simply establishing physical layer connectivity proved to be a time consuming and difficult task. Initially successful end-to-end connectivity was thought to be constrained by random hardware failures in the HDR earth stations or HIPPI/SONET gateways. Later, very close to the end of the experimental window, the physical layer connectivity problem was traced to the HIPPI/SONET gateway's SONET incompatibility with the HDR. Section 9.1 identifies this incompatibility and a subsequent workaround solution.

While this resulted in a great deal of valuable spacecraft time being lost, enough successful end-to-end periods were salvaged to verify the TCP paradigm over the spacecraft. As expected when sound SONET physical layer connectivity existed between the two supercomputers at NCAR and OSC, the TCP performance paradigm held up.

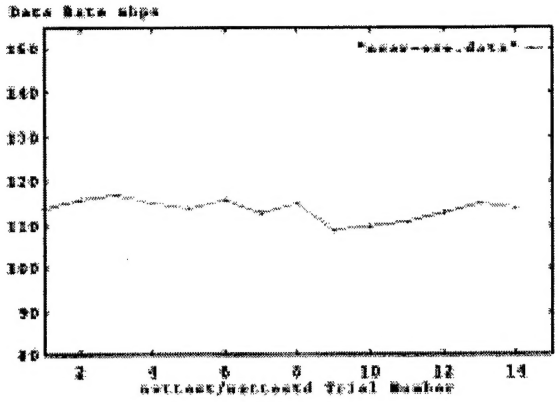

Figure 19: NCAR-OSC nettest/nettestd performance over ACTS.

Figure 19 depicts TCP performance via the hybrid network connecting the NCAR Cray J-916 and the OSC Cray Y-MP8 validating the TCP Performance Extensions outlined in RFC 1323 [1]. These parameters were conveyed to the researchers for incorporation into the coupled model-PVM applications.

8 HIPPI <-> ATM Performance

Due to time constraints HIPPI to ATM performance over the hybrid was not validated. However with the exception of the Maximum Transmission Unit (MTU) and hence the Maximum Segment Size (MSS) the performance parameters were the same as those between the Crays over the hybrid. Optimum performance is obtained in any environment by transmitting as large a packet (IP) as practicable. An MTU of 64 Kbytes for the HIPPI physical layer was used for the Crays while one of 9188 bytes (Classical IP AAL5 encapsulation) was used by ATM for the video conferencing and collaborative

workstations.

ATM to HIPPI performance was validated from the NCAR Cray J-916 to the NCAR SGI Onyx. The path extended from the Cray J-916, over HIPPI to the NetStar GigaRouter, to the SGI via ATM. The model's visualization output, a HIPPI stream from the Cray, was converted directly to ATM over SONET by the NetStar GigaRouter, obviating the need for the LANL HIPPI/SONET gateway.

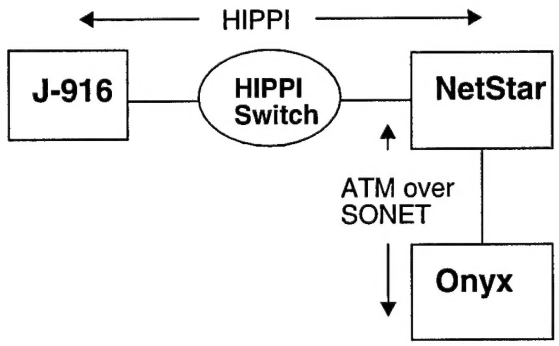

Figure 20: HIPPI to ATM over SONET via GigaRouter.

Dynamic MTU discovery on both machines facilitated the correct MTU for the data link layer used (9188 bytes) based on each machine's installed configuration.

The data in Figure 21 examines the nettest/nettestd performance between the Cray J-916 and the SGI Onyx. The noticeably smaller window sizes reflect low latency in the terrestrial fiber segment at the NCAR end of the hybrid.

The performance in both directions, alternating client and server is consistent and shows a marked performance increase with a larger window size. The SGI nettest/nettestd commands do not specify a -s option. This is due to the fact that a SGI workstation will invoke the appropriate window shift when the -b option signals a socket buffer size greater that the default window size set in the UNIX kernel. The Cray however requires the -s option, but not in this case, as the window size does not exceed the default. Large socket buffers are not required due to a trivial bandwidth-delay product.

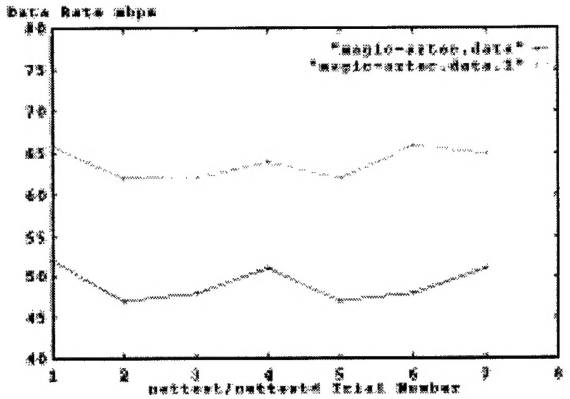

Figure 21: SGI-Cray nettest/nettestd performance over terrestrial GigaRouter segment.

It was not possible however to tune the SGI workstations for the larger window sizes required over the hybrid due to a kernel buffer limitation discovered in IRIX 5.2. A maximum kernel buffer space of 512 K bytes prevented the use of the optimum window sizes for the collaborative workstations.

9.1 HIPPI/SONET gateway - HDR SONET Section Integration

A major objective of the project was to investigate the interoperation of different physical layer technologies into the hybrid architecture. Prototype HIPPI/SONET gateways were integrated to serialize the parallel HIPPI datastreams for mapping to the SONET SPE of the ACTS HDR's. A great deal of time was spent researching problems establishing loopback as well as end-to-end connectivity through this device. The result was a loss of scheduled time on the spacecraft due to earth segment unavailability. This had an overall adverse impact on the progress of the entire experiment. Project milestones were not reached, causing a slide to the right of the experimental time line.

It was proposed that this phenomenon was the result of synchronization problems between the gateway and the HDR. Numerous earth station components were replaced in an effort to locate this inconsistency. Toward the end of the project BBN, LANL, NCAR and OSC engineers determined that the cause was a continuous cyclic reset between the two components. A procedure was developed to circumvent what appeared to be a State Machine

problem with the SONET section of the HDR .

Upon acquiring the spacecraft the HIPPI/SONET gateways required reset in such a manner as to establish synchronization during a perceived time window. Timing must be set to source on **BOTH** HIPPI/SONET gateways, each recovering timing from the other. The sequential order of synchronization was critical, once found it consistently established immediate end-to-end connectivity between the two sites. The procedure was used with complete success for the duration of the experiment. Regrettably this discovery was made late in the life cycle of the experiment and could not recover time lost.

The procedure is outlined as follows:

1. Terminate all traffic over spacecraft link.
2. Turn off HIPPI/SONET gateway and FIFO at first site.
3. Leave HIPPI/SONET gateway and FIFO down at first site during power-off of the HIPPI/SONET gateway and FIFO at second site.
4. Restart FIFO at second site, and restart HIPPI/SONET gateway after FIFO.
5. Restart FIFO at first site, and restart HIPPI/SONET gateway after FIFO.
6. Do not attempt to transmit ANY data traffic until verification of HDR state and spacecraft (TDMA) acquisition at both sites.
7. If state of both HDR's not OK go back to step 2, and begin again.
8. If state of both HDR's is OK initiate data traffic and evaluate link performance..

9.2 Rodent Damage to Outside Earth Station Components

Spacecraft time was lost to earth segment unavailability as the result of rodent damage to the power cabling for the Low Noise Amplifier (LNA) on the HDR terminal. The cabling connecting the outside LNA to the interior power supply was completely dissected by rodents. While it took a fair amount of time and effort to locate the source of earth station inoperability, repair was simple and complete functionality was restored quickly. Outside components were protected as well as practicable, however more intense hardening of facilities and routine, periodic inspection for such damage in future HDR implementations may help avoid such downtime.

10 Conclusion

In spite of the difficulties, the goals of the experiment were met. Adequate performance was obtained from a terrestrial/satellite hybrid network architecture to support interactive data communications between high performance end systems here-to-fore requiring low latency, high bandwidth interconnections.

New ideas were developed and proven which revitalized existing technologies. The venerable technologies of geostationary satellites were enhanced to provide high capacity, fiber optic quality communications in a new arena. Mature and proven by its functionality on the Internet, the classical TCP/IP protocol suite was successfully adapted to operate over modern, high capacity physical layer technologies. Successful integration of dissimilar physical layer architectures was also achieved, permitting interoperability between legacy systems and modern fiber-based communication technologies.

Inexpensive and reliable high bandwidth satellite communications systems like ACTS could increase the utility of high performance computing. Such systems could facilitate real-time, collaborative efforts in scientific research between non-collocated researchers anywhere in the world at any time without reliance on terrestrial systems. Earth and atmospheric scientists could access supercomputing resources immediately from sites in Lesser Developed Countries or remote areas not served by modern high capacity data communications, enabling them to advance the state of scientific discovery sooner and at less expense.

Future deployment and availability of wideband satellite communications systems will be a critical enabling technology of the emerging Broadband Integrated Services Digital Network (B-ISDN). These systems will provide cost effective, high capacity communications for the integrated voice, data and imaging applications which will make up the National Information Infrastructure (NII) and Global Information Infrastructure (GII). The presence of systems like ACTS will accelerate the successful deployment of these backbones.

References

[1] V. Jacobson, R. Braden and D. Borman. TCP Extensions for High Performance, May 1992. Request for Comments 1323.

[2] A. Geist, A. Beguelin, J. Donagarra, W. Jiang, R. Manchek and V. Sunderam. "PVM 3.1 Users Guide and Reference Manual'" Tech. Rep. ORNL/ TM-12187, Oak Ridge National Laboratory, May 1993.

[3] J. Postel. Transmission Control Protocol, September 1981. Request for Comments 793.

[4] J. Merrill. Pitch and Catch, National Center for Atmospheric Research. June 1995.

[5] B. Irwin. Modified nettest/nettestd utilities, National Center for Atmospheric Research. July 1995.

[6] D. Comer. "Internetworking with TCP/IP Volume 1: Principles, Protocols and Architecture,

Third Edition" Prentice Hall, Englewood Cliffs, NJ, 1995.

[7] D. Brooks T. Carrozzi P. Dowd, F. Lopez, F. Pellegrino and S. Srinidhi. "ATM Based Geographically Distributed Computing over ACTS" NASA Lewis Research Center, Cleveland, OH. 1995.